Eurosibs: Towards robust measurement of infant neurocognitive predictors of autism across Europe

Jones, E.J.H.*, Mason, L.*, Begum Ali, J., van den Boomen, C., Braukmann, R., Cauvet, E., Demurie, E., Hessels, R.S., Ward, E.K., Hunnius, S., Bolte, S., Tomalski, P., Kemner, C., Warreyn, P., Roeyers, H., Buitelaar, J., Falck-Ytter, T., Charman. T., Johnson, M.H., & the EuroSibs Team

Abstract

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that affects social communication skills and flexible behaviour. Developing new treatment approaches for ASD requires early identification of the factors that influence later behavioural outcomes. One fruitful research paradigm has been the prospective study of infants with a first degree relative with ASD, who have around a 20% likelihood of developing ASD themselves. Early findings have identified a range of candidate neurocognitive markers for later ASD such as delayed attention shifting or neural responses to faces, but given the early stage of the field most sample sizes are small and replication attempts remain rare. The Eurosibs consortium is a European multisite neurocognitive study of infants with an older sibling with ASD conducted across nine sites in five European countries. In this manuscript, we describe the selection and standardization of our common neurocognitive testing protocol. We report data quality assessments across sites, showing that neurocognitive measures hold great promise for cross-site consistency in diverse populations. We discuss our approach to ensuring robust data analysis pipelines and boosting future reproducibility. Finally, we summarise challenges and opportunities for future multi-site research efforts.

Key words: Infancy, neurocognitive, multisite, biomarker, eyetracking
Word Count: 11799

Introduction

Autism Spectrum Disorder is a neurodevelopmental disorder characterised by difficulties with social communication, and the presence of restricted/repetitive behaviours and sensory issues (DSM-5, APA, 2013). ASD affects around 1 to 2% of children and is associated with lifetime healthcare costs of up to £1.5 million in the UK (Buescher et al., 2014). Developing new ways to identify and support individuals with ASD to reach their full potential is thus critical. Indeed, early intervention can alter trajectories and improve outcomes (Dawson et al., 2010; Green et al., 2017; Pickles et al., 2016). However, although parents first report experiencing concerns in infancy or toddlerhood (Herlihy et al., 2013), the average age of diagnosis of ASD remains around 5 years in some countries (Brett et al., 2016). One issue is a lack of understanding of the earliest manifestations of ASD, which limits our ability to design novel early interventions and identify children whom they may benefit. Further, in order to optimise early identification and treatment we need to understand the mechanisms that underlie the emergence of ASD symptoms in early development.

Prospective longitudinal studies of infants with an older sibling with ASD ('infant siblings') allow us to study symptoms as they emerge (Jones et al., 2014; Szatmari et al., 2016). ASD is a highly heritable condition, with genetic contributions to phenotypic variation of 64-91% (Tick et al., 2016). Over the last decade, a number of prospective studies have followed infant siblings to the age of 3, when a diagnosis of ASD can be made with robust reliability (Woolfenden et al., 2012). Such studies have identified an ASD recurrence rate of around 20% (Ozonoff et al., 2011), with a further 20% of children experiencing sub-threshold symptoms or other developmental difficulties (Charman et al., 2017; Messinger et al., 2013). This research design is a powerful method of examining the pattern of infant development associated with a later diagnosis of ASD.

One of the major strengths of the infant sibling design (relative to retrospective approaches such as analysis of home videos or parent report) is the ability to assess neurocognitive development using infant-friendly methods such as eyetracking, electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS). Such techniques allow us to measure changes in brain and cognitive development that may occur before the clear emergence of behavioural signs of ASD. Indeed, early reports from infant sibling studies have suggested a number of promising avenues for further work (Jones et al., 2014; Szatzmari et al., 2016). For example, studies using EEG and eyetracking have indicated that infants with later ASD show reduced engagement with faces by 6 months in some contexts (Jones et al., 2016; Jones & Klin, 2015; Chawarska et al., 2013; Shic et al., 2014; Klin, Schultz & Jones, 2015), whilst overt social behaviour is broadly still typical (Ozonoff et al., 2010). Novel methodologies like fNIRS suggest there may be reduced specialisation of core social brain regions in infants with later ASD (Lloyd Fox et al., 2017). Atypicalities are not limited to the social domain. Slower disengagement of visual attention around the first birthday has emerged as a consistent finding across multiple reports (Elison et al., 2013; Elsabbagh et al., 2013; Zwaigenbaum et al., 2005). Disruptions have been observed in very low-level processes such as the strength of the pupil response to light measured with eyetracking (Nystrom et al., 2018). Strengths have also been identified; for example, eyetracking measures suggest that infants with later ASD may be better at identifying a target amongst distractors (Gliga et al., 2015; Cheung et al., 2016). A range of alterations in EEG power and connectivity have also been identified from early infancy, broadly suggesting altered trajectories of functional brain development (e.g. Bosl et al., 2011; Gabard-Durnam et al., 2015; Righi et al., 2014; Tierney et al., 2012; Levin et al., 2017; Orekhova et al., 2014). Early evidence thus points to the relevance of both socially-relevant

brain processes and early domain-general differences as critical areas of study in infant sibling designs.

Whilst promising, progress in understanding early functional brain development in ASD has been hampered by relatively small sample sizes (as in developmental psychology more broadly, though the ManyBabies project is a notable exception here; Frank et al., 2018). Further, there have been few attempts to examine the robustness of effects across contexts and cultures. Thus, there is a critical need for large-scale data pooling efforts and the integration of data from infants with a wide range of demographic and cultural backgrounds. One advance in yielding robust and clinically-relevant insights into infant siblings at the behavioural level has been the efforts of the Baby Sibs Research Consortium (BSRC) to pool behavioural, clinical and questionnaire data across multiple predominately North American groups. This data pooling effort has led to papers on the ASD recurrence rate (Ozonoff et al., 2011); prevalence of the Broader ASD Phenotype (Messinger et al., 2013); clinically-relevant difficulties beyond ASD experienced by infant sibs (Charman et al., 2017); sex differences in the expression of ASD symptoms (Messinger et al., 2015); and stability of diagnosis from 18 to 36 months (Ozonoff et al., 2015). The BSRC illustrates the power of pooling data across multiple sites in yielding insights into early development in ASD. However, since the BSRC focuses predominantly on pooling behavioural measures, the potential for multisite studies to answer broader questions about underlying neurocognitive mechanisms remains unexplored.

Large samples will allow us to ask more nuanced questions about the relations between different neurocognitive markers. For example, preliminary evidence from one relatively small cohort suggests that elevated disengagement times and altered gaze following make additive contributions to later ASD (Bedford et al., 2014); and that there may be sex differences in the relation between infant neurocognitive markers and later symptoms (Bedford et al., 2016). However, larger samples are needed to rigorously test these models. Further, larger samples will allow us to take seriously the mounting evidence that ASD does not reflect a unitary construct and may require stratification into more homogeneous etiological subgroups (Loth et al., 2015, 2017). Most studies to date have reported main effects of categorical diagnostic outcome , with far more limited data on individual-level prediction (though see Bussu et al., 2018). The latter is important not just for clinical utility, but also for determining the extent to which results represent a common path to ASD or are specific to a subgroup of individuals who may share a particular developmental etiology. Large prospective studies of infants with higher familial likelihood of developing ASD may allow us to subgroup types of ASD by developmental profiles, which may be the point at which varied etiologies are most clearly identified. Of note, we choose to use the term 'likelihood' rather than the more commonly used 'risk' to reflect the preferences of the autistic community (Fletcher-Watson et al., 2017).

Whilst data pooling efforts for behavioural measures are advanced, there are very few examples of attempts to prospectively collect neurocognitive data on a large scale. One exception is the Infant Brain Imaging Study (IBIS http://www.ibis-network.org/), a multisite US neuroimaging study which has produced a range of groundbreaking papers on structural brain development in infant siblings (e.g. Hazlett et al., 2017; Elison et al., 2013; Wolff et al., 2015). For example, data from the IBIS study suggests that ASD is associated with a pattern of early region-specific cortical hyper-expansion (Piven et al., 2017), and machine-learning approaches suggest that these observations have a degree of predictive validity at the individual level (Hazlett et al., 2017). These studies indicate the power of data pooling across multiple sites for generating insights into brain structure and function, as well as behaviour. However, MRI-based measures are expensive and can be very challenging to acquire in young infants. Their scalability as tools for early identification and their ability to identify features that could provide targets for intervention is thus a challenge, particularly as we

move to global health contexts. Further, it is very difficult to record MRI in awake infants. Since social interaction is a core domain of difficulty in ASD, measures of functional brain development that can be recorded in more naturalistic settings are also critical for understanding the disorder. Multisite studies of neurocognitive development in infants with a heightened likelihood of developing ASD tested across a range of cultures and contexts are necessary to move the field forward.

Here, we describe the formation of a common neurocognitive protocol that we have deployed across 9 sites in 5 European countries through the Eurosibs consortium. Eurosibs is a European multisite study involving investigators in the UK (Birkbeck College, London; King's College London, University of Cambridge), Sweden (Karolinska Institute, Uppsala University), Belgium (Ghent University), Poland (University of Warsaw), and the Netherlands (Radboud University Nijmegen, Utrecht University). We briefly outline how we selected and implemented a rigorously standardised experimental protocol at each site; how we standardised lab practices to maximise comparability; and how we designed analytic pipelines for pooling data across different collection systems. We present example data quality metrics from the subset of network sites who currently have sufficient sample sizes for each measure to illustrate the strengths of our approach to harmonisation. At this stage we refer to sites anonymously with letter codes to prevent premature consideration of specific cultural or linguistic factors that could affect our data because the study was not designed for this purpose; rather, we focus on identifying where measures appear consistent across sites. We further present experimental data from one example paradigm (the gap-overlap task) to test our ability to robustly measure core neurocognitive metrics across sites. We focus on this paradigm because slowed disengagement at 12- to 14-months has been related to later autism across three small previous studies (Elsabbagh et al., 2013; Elison et al., 2013; Zwaigenbaum et al., 2005); thus, it is a core target for our research program. Finally, we discuss our approach to minimising publication bias and the 'file drawer' problem, pitfalls and strengths, and our plans for building on our consortium work in the future. We hope that our study will provide an initial roadmap for other investigators interested in conducting multisite neurocognitive studies of early human development.

Methods

Participant information

Infants in the High-Likelihood group (HL) had an older sibling with a community clinical diagnosis of Autism Spectrum Disorder (ASD) as confirmed by clinician judgment. Infants in the Low-Likelihood (LL) group had at least one older sibling with typical development (as reported by the parent) and no first-degree relatives with ASD. Further information about clinical ascertainment can be found in S1.0. Infants were primarily enrolled at 5 or 10 months; a few infants were enrolled at 14 months. At each age point, the preferred testing window was +/-1 month from the relevant birthday; if this was not possible, we allowed testing up to +2 months to minimise data loss. In the present manuscript, we report available data metrics from a subset of infants tested at 5, 10 and 14 months. Some sites did not collect EEG or eyetracking data in infancy and hence are not included in the present report. Other sites only collected particular data streams at particular time-points (e.g. Site E did not collect infant EEG from Eurosibs tasks; site D only collected EEG at 5 months). Because data collection is ongoing, we have included all data uploaded to the central database and subjected to quality control assessment before February 2018 in this preliminary report. This data is thus intended to illustrate our protocol and procedures, and not to represent a finalised report on the cohort. Thus, all metrics are presented collapsed across likelihood group to avoid compromising future analysis plans, which are preregistered and embargoed until our defined data freeze points (see Discussion). Measures analysed, sample sizes and likelihood, gender and age profiles of samples at each site are shown in Tables 1a and b. Of note, there were significant differences between sites in the proportion of high and low likelihood infants in the sample (X^2= 13.6, p = 0.009), with site E having a relatively even balance (due to study design) whilst most sites had a majority of high likelihood infants. There were also significant age differences between sites (max. ηp^2 = 0.19), although numerically mean differences were small (Table 1b).

<<INSERT TABLES 1a and 1b ABOUT HERE>>

Ethics

Protocols were approved by the relevant committee at each site and were conducted in accordance with the Declaration of Helsinki and the American Psychological Association. Broadly, each study was conducted guided by a consistent set of principles. These included placing the infant's wellbeing at the centre of our focus at all times. Infants were always with their parent or caregiver who provided informed consent before the study began, and testing could be stopped or interrupted and rescheduled at any time if the infant became fussy or the caregiver did not want to continue. Protocols were designed to maximise infant comfort, and be interesting and engaging. Data protection and confidentiality are paramount, which shapes our approach to data sharing within and outside the consortium. Briefly, personal data is stored securely at each individual site; pooled data is pseudonymised and housed in a password-protected encrypted centralised database. Access is fully audited, and to ensure data security is governed by a management structure that includes cross-site Quality Control groups, modality-specific Core Analysis groups, and the Eurosibs management board (see Discussion and S1.2).

Protocol

Our full phenotypic protocol is shown in Table S1. This includes a range of phenotypic questionnaires that were translated by publishers (or translation services for in-house

measures) and administered in the native language at each site. Where possible, a full adaptation process with translation and back-translation consulted by the authors was undertaken (see Bolte et al., 2016 for an extended discussion of access to standardised measures in early autism research in Europe). Metrics were then harmonised within a central database. We have depicted cross-site data from two selected measures that are usually interpreted with respect to US norms; the parent-report Vineland (Sparrow, Bella, Cicchetti, Harrison & Doll, 1984) and the examiner-administered Mullen Scales of Early Learning (MSEL; Mullen, 1995). These were selected for compatibility with the BSRC consortium to allow future global data pooling; there were no available behavioural instruments normed in all the countries in our consortium. Site A used the US versions of the Mullen and Vineland. Sites B, C, and E used informal translations of the MSEL. Site D used the US version of the Mullen (including norms), with items requiring the examiner to speak translated informally into the local language by experienced clinicians and members of the testing team. Site E used a translated 72-item form of the Vineland survey questionnaire by E.M. Scholte & I.A. Van Berckelaer-Onnes, 2008 (based on S.S. Sparrow, A.S. Carter, & D.V Cicchetti) published and distributed by PITS B.V. Leiden, The Netherlands, with permission of the authors. At Sites B, C, and D the Vineland was administered as an interview by a trained clinician or researcher at the infant timepoints due to lack of availability of an officially translated version. Sites B and C did this based on a manual translated by S. Breider en A. de Bildt (2014, Leiden University); Site D used an in-house translation. The interview version of the Vineland was generally done in person during the visit, with a few exceptions conducted over the phone. Of note, all sites that used informal translations purchased the equivalent number of copies of the original forms from the publisher.

Neurocognitive measures were selected on the basis of preliminary evidence that they provided robust data quality, and that they were likely to be informative within the infant sib design (see below). We have demonstrated that many of our measures show good to high test-retest reliability (Cousijn et al., 2017; Hessels et al., 2016; Wass & Smith, 2014). We selected tasks that each yielded multiple informative metrics for analysis (rather than more 'cognitive' tasks that might provide one dependent variable). This allowed us to collect a denser quantity of information within a shorter time-window. In addition, some infants participated in an MRI and NIRS protocol at 5 months; because this was a relatively small proportion of the cohort, these datasets are not discussed here. Sites were required to administer tasks in a certain order within a testing modality (e.g. in the EEG session, children were assigned to view the social and non-social videos in a set order, specified in the scripting framework), but were free to adopt more flexibility in terms of when each modality was collected during the visit. This was essential since (for example) some families travelled longer distances than others, and so testing was split across more than one visit in some cases. However, where possible sites collected the EEG and eyetracking sessions after the child had napped (or when they first came into the lab) to minimise fatigue. We collected information about order on a Behavioural Observation Sheet.

Mullen Scales of Early Learning: The Mullen Scales for Early Learning (Mullen, 1995) are designed to measure cognitive abilities from birth to 68 months of age across 5 domains (Visual Reception [VR], Fine Motor [FM], Receptive Language [RL], Expressive Language [EL], and Gross Motor [GM]). Standard scores are calculated based on US norms. For harmonisation, we created a set of age-specific demonstration videos with scoring (in English) that were used in initial training at every site. All primary testers in charge at each site speak fluent English, as is common in European academia.

We hosted a series of in-person meetings, webinars and phone calls with core team members in which we went through the manual on an item by item basis and discussed administration and scoring. We discussed any queries arising during the course of the study at our monthly cross-site phone calls. Scores were reviewed for impossible values or missing items and corrected. Where subsequent training was required at an individual site (e.g. because a new PhD student joined the team), the primary testers would select examples of 'good administration' in the native language to complement the cross site training video. The new tester would watch these plus the original training videos; they would then watch live administrations at the site in question; read the manual and translation thoroughly; and then be observed for their first minimum three administrations with feedback. Administration would continue to be monitored by primary testers throughout the study.

Vineland Adaptive Behaviour Scale: The Vineland Adaptive Behaviours Scale-II (Sparrow et al., 2005; Sparrow et al., 1984) caregiver/ parent survey is a questionnaire that assesses adaptive behaviours in individuals from birth through 90 years of age across four key domains (communication, socialization, daily living skills, motor skills). Possible scores are based on how often the child performs the relevant activity, and include 2 (usually), 1 (sometimes or partially), 0 (never), NO (no opportunity), or DK (do not know). Again, standard scores are computed based on US norms. Sites A and E administered the questionnaire format because they did not have the resources or time to administer the Vineland as an interview at the infant testing points. The primary focus of our study was the neurocognitive measures, and so we chose to prioritise those batteries for the time the infant and parent were in the lab. Further, Site A had been using the Vineland questionnaire format for several years, meaning that continuing to use that measure could allow data pooling (Bussu et al., 2018). To ensure fidelity, in addition to the instructions to parents provided on the official version of the survey, sites provided parents with clear instructions about where to start and stop for each section. Irrelevant sections of the form (e.g. for younger or older infants) were crossed out. Forms were checked by trained research assistants for errors when the questionnaire was returned (e.g. not following the stopping rules correctly, missing items) and the parent was asked about these items in the visit. Due to lack of availability of an official translation for the questionnaire version, at Sites B, C and D administration was done in interview format by a trained clinician or research assistant. Sites entered data through the Vineland Assist, which provides a further level of automated data quality checking (e.g. whether ceiling has been correctly identified). Finally, we had a centralised quality control (QC) process that included double-entry and checking and correcting for impossible values and missing data.

Spontaneous EEG: Our core EEG task involved presentation of naturalistic videos with social and nonsocial content (Jones et al., 2015). The video stimuli were made up of women singing nursery rhymes (Social; see S1.3 for further details) or toys spinning (Non-social). The Non-social video comprised of 6 consecutive toys spinning, each spinning toy was presented for 10 seconds. The duration of each video condition (Social vs Non-social) was 60 seconds, and each condition was presented a maximum of 3 times, so infants received a total of 3 minutes per condition. The order of video presentation (Social or Non-social presentation first) was counterbalanced across participants. In the present paper we report metrics of data quality and completion rates from our EEG protocol.

Eyetracking: Four key eyetracking tasks were included in the full protocol (see S1.4 for further details). In the present manuscript, we present metrics of data quality taken across the battery. In addition, we report the core metrics extracted from the Gap task as an example of

our full processing pipeline. The Gap task (adapted from Johnson, Posner & Rothbart, 1991, Elsabbagh et al., 2013, Landry & Bryson, 2004) was selected because attention-shifting has shown predictive value for later ASD across three samples (Elsabbagh et al., 2013; Zwaigenbaum et al., 2005; Elison et al., 2013). Our operationalisation is a gaze-contingent paradigm that measures visual-attention shifting from a central to a peripheral stimulus in one of three conditions; i) Gap, in which the central stimulus disappears 200ms before the appearance of the peripheral target; ii) Baseline, in which the central stimulus disappears simultaneously with the appearance of the peripheral target; iii) Overlap, in which the central stimulus remains on screen during peripheral target presentation. Latency to shift attention to the peripheral stimulus in the Baseline vs Overlap conditions (disengagement) and in the Gap versus Baseline condition (facilitation) are key derived variables.

Recording systems (see S1.5 for detailed descriptions per site)
We wanted to align our protocols without requiring the purchasing of identical recording hardware at each site. This makes the study more cost-effective and easier to roll out on a larger scale in the future; builds on existing expertise within lab teams (rather than requiring them to learn new approaches); and provides a test of the generalizability of our findings across differences in hardware to which we expect them to be robust. However, it makes standardisation more technically challenging, and can raise the likelihood of site differences. To facilitate this, we developed a stimulus presentation framework, TaskEngine (https://sites.google.com/site/taskenginedoc/) to optimise the data quality and efficiency of acquisition of EEG and eye tracking data in multi-site studies. Briefly, this offers a number of advantages: 1) experimental stimuli were specified in centimetres and transformed to pixels at runtime by the framework. This ensured that the size and position of stimuli was constant for different screen sizes and screen resolutions across sites; 2) a unified system of event-marking ensured compatibility between different EEG and eye tracking hardware. Events were defined for each experimental task, and passed to the framework which then sent them onward over whichever communication channel was in use (e.g. TCP/IP connection for EGI Netstation, serial port for Biosemi); 3) eye tracking data were acquired, managed and saved to disk by the framework. Timestamped eye tracking event markers were integrated into this data stream, ensuring that experimental events were identified by discrete points in time, rather than samples (since sampling rates differed between sites); 4) the framework produces a rich collection of QC and analysis scripts, exploiting the standardised data format of the stimulus presentation scripts. QC reports allow cross-site comparisons of data quality (see example in Figure S1).

Infant behaviour management protocols
Principles: Infant testing requires not only standardisation of equipment and experimental tasks, but also the broader environmental context including examiner behaviour. This is commonly neglected in the literature, but it could contribute critical variance across sites. To facilitate harmonisation, we held a week-long in-person meeting in London to discuss and agree on a standard series of lab protocols. The broad principles were applicable across the whole testing session, but some of the more detailed attention-getting protocols described below were only available during neurocognitive testing. Our protocols were built on the principles of standardization with flexibility. In infant testing it is essential to be responsive to the needs of the individual infant, because otherwise data quality will be poor. Thus, we agreed a standard set of responses and priority order to infant fussiness; a standard set of instructions to parents; and a standard set of strategies for experimenters to use during testing that were included into Standardised Operating Procedures employed at all sites. Fuller details of testing practices for EEG and eyetracking are in the Supplementary Materials

(S1.1). Briefly, fussiness is defined as excessive motion, negative affect and avoidance behaviour that indicates the child is not enjoying participating in the experiment. Significant fussiness typically leads to poor quality or missing data. To maximise cross-site standardisation in responses to fussiness, we agreed a hierarchy of responses to maximise both data yield and participant comfort. Before the experiment, researchers asked parents to maximise their baby's comfort by ensuring they were warm, fed, changed and seated comfortably. During the experiment, if a baby showed signs of fussiness (e.g. began to move more, show negative facial expressions, turn away from the screen) experimenters first addressed possible boredom by playing non-social 'attention-getters' through speakers. The presentation of these was automatically recorded in the data file, facilitating later comparison of the use of these strategies across sites. If that did not work, the examiner spoke to the baby; a manual code could insert this into the datafile. If fussiness continued, the parent was instructed to try (in this order) cuddling; holding hands; give baby something boring to hold (like a plastic teething ring); give baby a pacifier or snack; if all that did not work, a break was taken. If the parent wished to try again after the baby had calmed down, the experiment was resumed.

Quality control

Ongoing quality control was evaluated through monthly phone calls with site representatives during which data quality, testing practices and other issues were discussed. Further, a designated researcher with expertise in infant EEG and/or eyetracking supervised ongoing quality control checks of data as it was collected at each site. After collection, each data analysis pipeline includes further extensive quality control assessment. These will be analysis specific, but for example in EEG include rejection of segments where infants did not attend to stimuli, and segments containing artifact (e.g. motion, poor electrode contact).

Each site collected session-level data on a Behavioural Observation Sheet. This included observations like changes in EEG cap position, infant state, any interference from parent or examiner, or any contextual variables that might affect data interpretation (like presence of an additional child). Further, the examiner reported their opinion of data validity for each modality. Specifically, data was marked as valid (can proceed to the next stage of processing and analysis); questionable (video of the session should be evaluated); or invalid (data should not progress to the next analytic stage). Data marked invalid was further assigned a qualitative category to denote the reason for invalidity. These were reasons to do with the child (e.g. would not wear the EEG cap, inattention, fussiness); reasons to do with the examiner (e.g. experimenter error); reasons to do with the parent (e.g. refused to participate, asked for session to stop, constantly interfered); technical error (e.g. computer crashed); or other. These categories are critical to our ability to evaluate the reasons for data loss. Data marked valid or considered valid after review would move to the next step of the processing stream (at which further validity and inclusion judgements will be made).

Eye tracking data was acquired by the TaskEngine stimulus presentation framework, tagged with anonymised metadata and saved to disk. Offline preprocessing was performed in a custom-written pipeline in the following stages: 1) Data was loaded and IDs, age point and site metadata recorded; 2) The reference frame of the eye tracking data was transformed to harmonise the effects of screen size across sites. For sites with larger screens, stimuli were presented in a central 'virtual window' within the main screen. In these cases, the reference frame of the data was relative to the physical screen, this transformation made it relative to the virtual window; 3) Data was segmented by task and saved to disk; 4) Data quality metrics were extracted for all sessions and all tasks.

Analysis approach

In the present manuscript, we report cross-site comparisons for our selected behavioural measures (Mullen Scales of Early Learning, Vineland Adaptive Behavior Scales). We compare online data quality assessments for EEG data across sites. We then present an in-depth evaluation of cross-site data quality from our eyetracking battery. Briefly, the quality control metrics that we extract are: 1) Proportion of lost data samples (due to blinks, looking away, or the eye tracker failing to detect the eyes); 2) Flicker ratio – the proportion of pairs of adjacent samples that are both either present or absent capturing the degree to which the contact with the eyetracker fluctuates on a sample by sample basis; 3) Precision, calculated as the root mean square of the Euclidean distance between adjacent samples (in degrees) during fixations (defined as a series of adjacent samples with a velocity <1SD from the mean); 4) Accuracy, calculated as the mean offset of gaze from the centre of post-hoc calibration stimuli in degrees, separately for the *x* and *y* axes; 5) Mean distance of the infant's eyes from the screen (distance along z-axis of track-box), and from the centre of the track-box (3D) in mm; 6) Standard deviation of the distance of the infant's eyes from the screen and centre of the track-box in mm.

Finally, we analysed the gap/overlap task and extracted metrics related to data quality and its effect on an experimental task: 1) Number of valid trials; 2) Mean and SD of saccadic reaction time (SRT) from the *baseline* condition; 3) Mean and SD of the disengagement effect (overlap condition - baseline condition). Full data processing details are given in section S1.6 of Supplemental Material. Briefly, through automated scripts each trial was checked for data quality and correct behaviour on the part of the infant (moving directly from the central to the peripheral stimulus). Invalid trials were excluded, and the SRT to shift from the central to the peripheral stimulus was averaged across all remaining trials. All metrics were analysed with one-way ANOVA, with site as the independent factor, separately for each time point. Analyses were not corrected for multiple comparisons to avoid being overly conservative in detecting effects of site.

Results

Sites were broadly comparable in the maternal education status of enrolled families (Table 2). A chi-squared test on the number of parents with primary/ secondary (education to 16 years) vs tertiary education revealed marginally significant site effects ($\chi^2 = 6.52$, $p = 0.09$), though given the size of the sample the differences were not pronounced. Broadly, Table 2 indicates that parents had predominantly high levels of education. Of note, here we have operationalised socio-economic status as maternal education because this is a common metric to use in the US, and raw income bands are challenging to harmonise across countries. However, access to higher education also varies in cost and population take-up across Europe and may be influenced by factors distinct from those that influence higher education access in the US. We are actively exploring other options, such as computing ratios of incomes relative to poverty thresholds in each country.

<< INSERT TABLE 2 ABOUT HERE>>

Behavioural Assessment

Scores on the Mullen Scales for Early Learning and the Vineland Adaptive Behaviour Scales both domain standard scores computed on the basis of US norms) were analysed using multivariate ANOVA with Site as an independent variable. Statistics are reported in Table 3 (colour coded by effect size such that darker colours represent higher effect sizes), and data is visualised in SM Fig S2. Broadly, the pattern of data indicates substantial site effects on

Mullen scores. Site effects tended to be stronger for language measures (average ηp^2 across ages and RL, EL of 0.22; vs average ηp^2 across ages for motor [FM, GM] domains of .09), and stronger at older age points (average ηp^2 of 0.10 at 5m; 0.17 at 10m; 0.21 at 14m).

<< INSERT TABLE 3 ABOUT HERE>>

There were also site effects on Vineland scores, though they were generally smaller than Mullen scores (Table 4; visualised in SM Fig S2). Again, effect sizes were stronger at older age-points (mean ηp^2 of 0.04 at 5 months, 0.09 at 10 months and 0.11 at 14 months). In contrast to the Mullen, site effects were relatively stronger for daily living (mean ηp^2 across age 0.15) and motor domains (mean 0.11), and weaker for communication (mean 0.02) and socialisation (mean 0.03).

<<INSERT TABLE 4 ABOUT HERE>>

However, individual differences appeared equally stable over time at each site. Specifically, we conducted a series of ANCOVAs with 5-month Mullen domain scores as the covariates, site as an independent variable, and 10-month Mullen domain scores as the outcome variables (Table 5). We specified the model to test the main effects of site, 5-month Mullen domain scores and their interaction on 10-month Mullen domain scores. A significant interaction term would suggest that the sites significantly differed in the degree to which 5-month Mullen scores predicted 10-month Mullen scores on that domain. Broadly, this was not the case; and in all cases 5-month Mullen domain scores significantly predicted 10m scores. The same was true from 10 to 14 months. The pattern of results was similar for Vineland scores with the exception of socialisation between 10 and 14 months, which showed less stability at Site D ($F(1,116) = 1.49$, $p=0.23$, $\eta p^2 = 0.013$) than Sites A ($F(1,58) = 18.3$, $p<0.001$, $\eta p^2 = 0.24$) and B ($F(1,116) = 1.49$, $p=0.23$, $\eta p^2 = 0.013$). Thus, although there were significant site differences in overall scores, there was broadly similar developmental continuity over time between sites.

<<INSERT TABLE 5 ABOUT HERE>>

EEG

Because data has not yet been fully cleaned or processed, here we report the number of videos watched across sites, and the distribution of in-person validity judgements. This provides valuable information as to whether sites were able to collect sufficient data for analysis on this measure. Sites were broadly comparable in the percentage of children who provided valid data out of all the children in which data collection was attempted (Table 6). A chi-squared analysis showed no difference in the proportion of children with valid data by site at 5 months ($\chi^2 = 2.06$, $p = 0.15$); 10 months ($\chi^2 = 2.65$, $p = 0.27$); or 14 months ($\chi^2 = 2.98$, $p = 0.23$). Within groups of children who watched at least one video (i.e. excluding children who refused to wear the sensor net/cap), there were significant site effects for trial numbers at 5 months ($F(1,146) = 7.13$, $p = 0.007$, $\eta p^2 = 0.047$) and 14 months ($F(2,140) = 3.47$, $p = 0.034$, $\eta p^2 = 0.048$) but with modest effect sizes; there were no significant effects at 10 months ($F(2,176) = 1.51$, $p = 0.23$, $\eta p^2 = 0.017$). The average number of one-minute segments watched was over 5 out of 6 at all age points and sites, which indicates that this measure was well tolerated and produced a low drop-out rate.

<<INSERT TABLE 6 ABOUT HERE>>

Eyetracking

Since data quality metrics were highly intercorrelated (Table S6) we ran a factor analysis with oblimin rotation on measures of post-hoc drift, proportion lost samples, flicker ratio, precision, and mean and variability of distance to the centre of both the screen and trackbox. Barlett's test of Sphericity was highly significant (X^2 = 6988.4, p < 0.001). Eigenvalues indicated that two components best explained the data (60.6% of variance) and Kaiser's sampling adequacy was good (0.63). The two extracted components (with intercorrelation of 0.13) roughly corresponded to Contact (with loadings of over .8 for post-hoc drift, proportion samples lost, and flicker ratio) and Position (with loadings > .75 for variability in distance from the screen and track box, and > .5 for mean distance; see Table S4). Full results for all quality control metrics are detailed in Table S5 and visualised in Figure 1 and Figure S3 of the Supplemental Material.

<< INSERT FIGURE 1 ABOUT HERE>>

A repeated-measures ANOVA on the two factors by site indicated that there were significant site effects at both 10 and 14 months (10 months: $F(4,317) = 13.58$, $p < 0.001$, $\eta p^2 = = 0.15$; 14 months: $F(4,304) = 11.43$, $p < 0.001$, $\eta p^2 = = 0.13$) and an interaction between the type of metric and site (10 months: $F(4,317) = 3.86$, $p = 0.004$, $\eta p^2 = 0.046$; 14 months: $F(4,304) = 3.31$, $p = 0.011$, $\eta p^2 = = 0.042$). The pattern of site effects indicated broadly that sites C and D had relatively better Contact than other sites {Contact at 10 months: (B>D,C), (A>D,C), E, D, C; at 14 months (B>C,D),E,A,C,D} and that site D had relatively more optimal Positioning {10 months (C, A, E,B>D),D; 14 months: A>B,C,D), (E<D),B,C,D}. There were no significant site effects or interactions with data quality at 5 months (*p*s>0.05).

Our most critical comparisons were on our experimental measure, the gap-overlap paradigm. The mean number of valid trials acquired on the gap-overlap paradigm differed between sites at 10 months ($F(335)=2.61$, $p=.035$, $\eta p^2 = 0.03$), with marginal significance at 5 months ($F(193)=2.92$, $p=.056$, $\eta p^2 = 0.03$) and not at 14 months ($F(322)=1.27$, $p=.28$, $\eta p^2 = 0.02$); broadly Site D had more valid trials than other sites. To understand the relation to data quality, we used ANCOVA with trial number as the dependent variable, the two measures of data quality as the predictor variables and included site as a between-subject variable as a main effect and in interaction with the metrics of data quality. As might be expected, this indicated a significant relation between better Contact (lower scores) and a greater number of valid trials at 10 months ($F(1, 315) = 29.25$, $p < 0.001$, $\eta p^2 = 0.089$) and 14 months ($F(1, 302) = 12.42$, $p = 0.001$, $\eta p^2 = 0.041$), with a significant interaction with Site at 10 months ($F(1, 315) = 3.19$, $p = 0.014$, $\eta p^2 = 0.041$) but not 14 months: $F(1, 302) = 1.22$, $p = 0.30$, $\eta p^2 = 0.017$)). This reflected stronger relations at Sites A and E (r^2s = 0.28 and 0.35 respectively) vs the other sites ($r^2 < 0.15$). Of note, the main effect of Site was not present after Contact was included in the model, indicating that this may explain the Site differences in trial numbers ($F(1,301) = 0.71$, $p = 0.59$). Positioning did not significantly relate to trial number at 10 months ($F(1, 315) = 0.85$, $p = 0.36$, $\eta p^2 = 0.003$) but better Positioning related to more valid trials at 14 months $F(1, 302) = 4.87$, $p = 0.028$, $\eta p^2 = 0.02$ and this varied by Site ($F(4,302) = 3.42$, $p = 0.009$, $\eta p^2 = 0.045$). This reflected strong relations at Site E ($r^2 = 0.31$) but weaker relations elsewhere (r^2s < 0.11). At 5 months, better Contact also related to a greater number of valid trials (Contact ($F(1, 176) = 24.0$, $p < 0.001$, $\eta p^2 = 0.13$); but Positioning did not $F(1, 176) = 0.042$, $p = 0.84$, $\eta p^2 < 0.001$), with no interaction with Site (*F*s

< 2.4, *p*s > 0.1, ηp^2 < 0.03). Thus, as would be expected, better Contact at all ages (less flicker, fewer samples lost, less post-hoc drift) and better Positioning at 14 months (where infants are more mobile and perhaps more likely to move around) were associated with higher numbers of valid trials. This is consistent with the overall pattern of site effects, in that site D had more valid trials and generally better Contact and Positioning.

The mean baseline saccadic reaction time (SRT) also differed between sites at all age points (*F*s>2.507, *p*s<.042, ηp^2 >.03). Broadly, reaction times were slowest at site D and significantly faster at site A (Table S2). ANCOVAs as above indicated that at 5 months, neither metric of data quality related to SRTs (Contact $F(1, 157) = 0.25$, $p = 0.62$, $\eta p^2 = 0.002$; $F(1, 157) = 0.52$, $p = 0.47$, $\eta p^2 = 0.004$). There was a significant relation between better Contact (lower scores) and shorter baseline SRT at 10 months ($F(1, 303) = 6.23$, $p = 0.013$, $\eta p^2 = 0.021$) and 14 months ($F(1, 291) = 5.92$, $p = 0.016$, $\eta p^2 = 0.021$), with no significant interaction with Site at either age (although the effect size at 10 months was bigger than the main effect; $F(1, 303) = 1.68$, $p = 0.16$, $\eta p^2 = 0.023$; 14 months: $F(1, 291) = .75$, $p = 0.56$, $\eta p^2 = 0.011$)). However, the main effects of Site remained significant (10 months: $F(1, 303) = 4.22$, $p = 0.002$, $\eta p^2 = 0.055$; 14 months $F(1, 304) = 7.07$, $p = 0.008$, $\eta p^2 = 0.047$) with larger effect sizes, indicating that variation in Contact did not account for site variation in SRT. Positioning did not significantly relate to baseline SRT at either age (10 months: $F(1, 303) = 1.25$, $p = 0.27$, $\eta p^2 = 0.004$; 14 months $F(1, 291) = 3.49$, $p = 0.08$, $\eta p^2 = 0.047$). Thus, despite efforts to minimise the effect of data quality on task dependent variables (see S1.6), better Contact (less flicker, fewer samples lost, less post-hoc drift) remained weakly associated with faster baseline RTs in older infants (explaining 2-3% of the variance).

Of note, simple correlations between Contact and shorter baseline SRT remained significant when number of valid trials was partialled out at 10 months ($r(301)=.13$, $p = 0.029$) and 14 months ($r(289)=.13$, $p = 0.049$), indicating that the relation is not mediated by the number of trials obtained. The individual-level relations between better Contact and shorter baseline RT can be contrasted with the relatively better Contact and *slower* reaction times at site D. Thus, site differences in SRTs are unlikely to be simply related to site differences in data quality. Nonetheless, covarying effects of Contact in future analysis involving basic SRTs could improve signal to noise ratio.

Finally, we examined the primary dependent variable of interest from the gap task, the disengagement score. Disengagement scores are the difference between the baseline and overlap scores, and have been previously associated with later autism (c.f. Elsabbagh et al., 2013). Critically, neither the mean disengagement scores, nor the standard deviation, significantly differed between sites (*F*s<2.456, *p*s>.089, ηp^2 <.03). Further, ANCOVAs as above revealed no significant relations with Contact (5 months: $F(1, 157) = 0.68$, $p = 0.41$, $\eta p^2 = 0.005$; 10 months: $F(1, 303) = 0.004$, $p = 0.95$, $\eta p^2 < 0.001$; 14 months $F(1, 291) = 0.937$, $p = 0.334$, $\eta p^2 = 0.003$) or Positioning (5 months: $F(1, 157) = 0.03$, $p = 0.86$, $\eta p^2 < 0.001$; $F(1, 303) = 0.377$, $p = 0.54$, $\eta p^2 = 0.001$; 14 months $F(1, 291) = 0.542$, $p = 0.462$, $\eta p^2 = 0.002$) at any age. Thus, whilst scores on individual conditions may be influenced by data quality metrics, our primary metric of interest "disengagement" is robust to these lower level differences (Figure 2).

<< INSERT FIGURE 2 ABOUT HERE>>

Finally, we examined the specificity of individual differences in eyetracking data quality and core metrics over time at each site, to parallel behavioural analyses. Specifically, we conducted a series of ANCOVAs with each 5-month eyetracking variable as the covariates,

site as an independent variable, and 10-month eyetracking as the outcome variable, with and without our two data quality factors as additional covariates (Table 6). We specified the model to test the main effects of site, 5-month eyetracking scores and their interaction on 10-month eyetracking scores. A significant interaction term would suggest that the sites significantly differed in the degree to which 5-month eyetracking scores predicted 10-month eyetracking scores on that domain. Broadly, this was not the case. Significant between-age stability from 5 to 10 and 10 to 14 months that did not vary by site was observed for baseline SRT; this remained significant after Contact was covaried (although this then revealed stronger developmental continuity at Site E). Interestingly, Contact itself showed some continuity between 5 and 10 and 10 and 14 months; the latter varied by site such that continuity was stronger at Sites A and E (the largest two sites). Disengagement showed significant stability between 10 and 14 months (and after data quality was covaried) but not earlier in development. Position did not show stability in individual differences.

<<INSERT TABLE 7 ABOUT HERE>>

Discussion

We report methods and harmonisation procedures for the Eurosibs multisite study of infants with high likelihood of developing ASD. The push towards 'big science' and the growing recognition of the need for more rigorously powered studies means that research groups are increasingly seeking to join together to build larger co-ordinated cohorts. This may be particularly important for prospective longitudinal studies of high likelihood populations, for which only a small proportion of infants may develop the clinical outcome of interest. However, multisite studies of young infants bring a unique set of challenges, perhaps particularly when using neurocognitive methods that involve advanced and often site-specific basic equipment. Our European context brings new challenges but also new opportunities because of the linguistic, cultural and contextual variance across sites. Testing whether our data is robust across substantial variance in these factors (including in experimental hardware) provides one way to triangulate evidence for the robust relation between predictor and outcome. Further, if we wish our methods to extend to clinical practice in future work we must consider how harmonisation of measures across diverse recording sites can be achieved cost-effectively. By sharing our strategy, we provide valuable information to others in the field about meeting the challenges of multisite infant research.

Behavioural measures

There were significant site effects on our observational behavioural measure (the Mullen Scales) that were greater at older age-points, and greatest for language scales. There are several potential explanations for this pattern. Firstly, it may be that a scale designed to measure acquisition of language in US infants does not work in the same way in a European context. Scores are compared to US norms that were generated from comparatively small samples, and it is likely that new country or language-specific norms are required. Indeed, norms differ between Dutch and US samples on the Bayley Scales of Infant and Toddler Development, a behavioural test comparable to the Mullen Scales (Steenis, Verhoeven, Hessen, & van Baar, 2015). Since official translations of our measures were largely unavailable, sites used in-house informal translations where necessary. This could also have added to cross-site variance in scores. Further, language development may proceed differently in different cultures; to tap comparable abilities items may need to be adjusted on a per-country basis (Pena, 2007). However, this poses other challenges for harmonisation and data pooling: adjusting scales based on how typically developing infants acquire language in

different cultures may not account for the fact that infants developing autism may have their own idiosyncratic acquisition patterns of word acquisition (Lazenby et al., 2016). Further, there were also heterogeneous language experiential profiles within some of our sites (particularly for site A and E), with exposure to additional languages very common. Using a measure specific to each infant's primary language exposure would have been prohibitive and would have injected significant additional variance. Within our study, some sites have datasets per age point that are larger than the original normative samples for some of the instruments we used. One analytic strategy may be to re-norm our cognitive scores within each site for later pooling (Mullen, 1995). Alternatively, we could explore patterns of data at the item level to see whether there are particular items that may be less specific to a particular language and more consistent across cultures (like features of communication); or use newly developed cross-linguistic tasks that take into account word frequency and age of acquisition for multiple languages (e.g. LITMUS-CLT, Haman et al., 2017). However, we think that the challenges of translating behavioural measures comparably (including those driven by the publishing industry; Bolte et al., 2016) is a further point in favour of neurocognitive measures that can better measure the more basic abilities that the child uses to learn whatever language to which they are exposed.

Second, it may be that different testing practices across labs contributed to variance, despite our efforts to harmonise administration. Within-US multisite studies do not appear to have experienced significant difficulties in achieving cross-site reliability (Olson, Gotham & Miller, 2012) but it may be that lab practices vary significantly more within Europe than within the US where there is likely greater mobility of the workforce between centres. Thus, it may be necessary to increase the intensity of cross-site harmonisation procedures in future efforts, but this requires considerable resources. During the course of a study, if site effects emerge it is not necessarily the case that renewed efforts should be made to further harmonise procedures. Adjusting testing protocols during the course of a prospective longitudinal study is unwise. Since we do not know at the outset which children will develop ASD, we cannot ensure their even distribution across testing approaches. Further, changing testing protocols between timepoints could significantly impact longitudinal continuity. Our analyses suggest that within-site stability of developmental progression across age points was consistent across the consortium, indicating that our behavioural measures are likely valid measures of individual differences in developmental trajectory despite mean differences across locations. Thus, we chose to maintain current testing protocols at each site and deal with cross-site differences in later analysis. Third, it may be that there was true heterogeneity in the behavioural development of infants at each site. Sites did not significantly differ in maternal education but there was heterogeneity in the proportion of high-likelihood infants within each sample. There is also likely as-yet unmeasured heterogeneity in ASD outcome between the cohorts. This may contribute to site effects in behavioural assessments. However, site effects on the parent-report Vineland measure were smaller, and more pronounced for motor skills and daily living than for communication skills. If site differences were due to true differences in behavioural profile, we would expect them to be consistent in domain profile across measures. Taken together, these observations indicate the challenges of using observational behavioural assessments of cognitive development in European samples, which are likely to themselves be less challenging than spreading research to other low and middle income countries.

Neurocognitive measures

Our neurocognitive measures showed promise in yielding relatively low initial attrition rates. For example, good quality EEG data was obtained from 70-100% of infants, with high comparability in attrition rates across sites. Given the EEG systems used ranged from 32

channel gel-based systems used at home to 128-electrode water-based electrodes in the lab, this similarity of acquisition rates is promising. Similarly, 90.1% of infants produced sufficient trials for analysis in the eyetracking gap-overlap paradigm. These relatively high rates of data inclusion provide critical evidence that neurocognitive measures are feasible for use in multisite infant studies, particularly since at each site these measures were embedded in much longer batteries. However, we did observe some site differences in low level parameters, such as the number of valid trials available for analysis. As for behavioural assessments, the range of possible explanations for this finding include true behavioural heterogeneity, differences in lab testing practice, and differences related to recording hardware (e.g. sampling rate differences).

In contrast to behavioural assessments, our neurocognitive batteries provide far richer contextual data to aid interpretation of site effects. For example, our analyses and correlation tables (Figure S3) suggest that the number of valid trials suitable for analysis is primarily affected by raw data quality acquired by the eyetracker/eyetracker 'contact', including spatial error (post-hoc drift, precision) and temporal error (lost samples, flicker) at 5 and 10 months. We believe that this is an indication that our processing procedures for the gap task are effective at excluding trials with low basic data quality. Further, site differences in the number of valid trials obtained were only significant at 10 months, with infants at Site D providing the most valid trials for analysis, and also displaying the best Contact and Positioning. When Contact alone was taken into account, site effects were no longer significant and only Contact related significantly to valid trial number at this age. Thus, we can interpret site effects in trial numbers as resulting primarily from differences in the fidelity with which the infant's eyes could be tracked, rather than how much they moved during the session. Tracking fidelity may vary with hardware factors such as the model of eyetracker used: Sites C and D (with relatively better data quality) both used the TX300 eyetracker run at 300Hz (the latest model). In contrast, site differences in baseline SRTs could not be explained by differences in Contact. At all three ages, better Contact was associated with shorter baseline SRTs. However, whilst site D had broadly better Contact than other sites, SRTs at Site D were generally *longer* than at other sites. Site effects on SRTs also remained significant with increased effect sizes when Contact was covaried. Positioning did not relate significantly to SRTs. Thus, site differences in SRTs could not be explained by differences in basic data capture or in infant motion but could reflect true differences in emerging infant characteristics that will be further explored when longitudinal data is available.

At 14 months, when infants are older and more mobile, the number of valid trials for analysis also become related to measures of infant Positioning (mean distance and variability within the trackbox or with respect to the screen). This potentially reflects inattention (e.g. see also Miller et al., 2016), such that infants who are less attentive move around more and are willing to spend less time engaging in the task. Further exploration of metrics of infant motion during eyetracking paradigms and their interaction with neurocognitive data may contribute to our understanding of early-emerging behavioural profiles associated with later autism and ADHD.

Critically, we did not observe significant site effects on disengagement scores (as measured by overlap-baseline saccadic reaction times), our primary outcome variable for this task (Elsabbagh et al., 2013). This is highly promising since it suggests that whilst we were not able to achieve total harmonisation of basic saccade metrics (as evidenced by differences in baseline SRTs), the site differences were systematic enough that we *were* able to record harmonious condition differences. Disengagement scores were entirely independent of our quality control metrics (Table S6) indicating that our analytic pipelines were robustly able to generate putative biomarkers that are robust to variation in equipment and associated data quality. The one remaining correlation was between disengagement score and distance to the

screen at 14 months, such that infants seated further away from the screen saccade more rapidly. This may reflect the fact that distance from the screen affects the position of the peripheral stimulus on the retina, which would mean that a saccade takes longer to arrive at the peripheral stimulus. Having a precise measurement of this variable allows us to correct for this factor on a trial by trial basis. The disengagement scores are our final measure of task performance, extracted from the final stage of processing, which suggests that all of the preceding stages of data acquisition and processing were successful in allowing us to pool infant task data across multiple testing sites. Taken together, this demonstration of the potential of our pipelines to generate neurocognitive biomarkers that are robust to variation in both equipment and testing approaches is critical to their future deployment in clinical trial contexts, where standardisation of equipment across sites is impossible.

Good scientific practice

We recognise emerging debates in Psychology and other fields concerning reproducibility, p-hacking, bias, and other questionable scientific practices. We have developed guidelines to maximise the quality of our scientific reporting. All measures were collected to Standardised Operating Procedures (available on request). All data analysis is accompanied by rigorous presentation of data quality metrics to ensure accurate interpretation. All analysis projects must be pre-specified on either a web-based resource such as Open Science Framework, or internally through our database. Any data access requires a pre-approved project form to discourage unreported analysis. All data analyses conducted will be reported, either as a publication or as an internal report that will be hosted on our website for other investigators to download should publication be difficult. At regular intervals we will publish accumulated summaries of these unpublished materials when they are considered sufficiently substantive. Datasets and analysis scripts used for published analyses will be archived centrally on the Eurosibs servers so that they can be readily accessed for replication work by other investigators. We also take a variety of active approaches to encourage replication and task-sharing, which has led to a number of collaborative manuscripts (Haartsen et al., in press; Bussu et al., 2018; Tye, Bussu et al., in review; Braukmann et al., 2018) and the use of our TaskEngine framework by over 11 research groups in 8 countries.

Data collection for our study is still ongoing, and thus primary output data is subject to an embargo period. However, once collection is complete and a period has elapsed to allow a primary data report from the consortium, our data will be made available to the research community through a limited (curated) open-access procedure. Considerable effort has gone into developing data sharing and access procedures that balance our responsibilities to both participants and the broader community. We recognise that data sharing is critically important to maximising the benefit of research; however, we also must consider the need to protect the confidentiality and personal health information of this sensitive group (who as minors do not consent for themselves, and for whom in some cases the diagnosis of ASD made through the research study is not necessarily accompanied by a community clinical diagnosis; thus, the infants and families themselves may not want to know that they meet criteria for ASD). In particular, our data has maximum value when data points are linked together (e.g. early EEG and later ASD outcome). This pseudonymised nature of the data makes open sharing particularly challenging because linked data-points increase in identifiability. Some measures require the use of video material, which being inherently identifiable requires even stricter governance. Currently, our consortium governs access to the data to ensure that end users confirm their compliance with all relevant data protection laws and have appropriate ethical permissions to perform relevant analyses. This operates via a project approval form that is considered by the Eurosibs Board, which consists of representatives from all involved sites. Projects are evaluated primarily on their consistency

with the ethical principles families agreed to when they signed up to the study, and for overlap with other ongoing projects (in which case collaborations are suggested). However, this procedure is reviewed on a rolling basis to optimise our value to the scientific community.

Lessons learned

During the creation of Eurosibs, we have encountered many challenges that provide important lessons for future work. First, the time commitment and resources required for successful cross-site harmonisation should not be underestimated. We held a residential week-long training and harmonisation meeting, hosted several lab exchanges and site visits between labs within the consortium, and maintained standards through monthly phone calls and bi-annual in-person meetings. We also created standardised operating procedures for all our measures. The success of our consortium was in large part dependent on this high level of motivation. Despite this investment, site differences remained. One critical factor in infant testing is how to operationalise behaviour management. During our week-long residential meeting with core testers, we had lengthy discussions about strategies to deal with fussiness that revealed significant differences in existing lab practices. To address this, we agreed a harmonised protocol that attempted to balance flexibility to the needs of the infant with standardisation. For example, if infants started to become fussy (with increasing motion and negative affect) experimenters used a sequence of approaches that began with the use of attention getters coded and recorded through the scripts, moved through social strategies that minimised effects on data quality like cuddling or hand holding, and ended with taking a break. Broadly, this resulted in relatively comparable data quantity across sites. For example, all sites collected a per-baby average of over 5/6 EEG videos. However, there were some significant site differences in these metrics (albeit with small effect sizes). These were not consistent across age groups: no one site collected consistently larger quantities of data. Further, these differences did not appear to vary consistently with data quality such that it did not appear that (for example) one site was persisting with testing despite poor data quality longer than another. Thus, we do not see clear evidence for significant systematic site differences in how long testers persisted with the battery. However, the existence of some site differences may suggest that there were perhaps smaller fluctuations in testing practice that had less systematic and pervasive effects on data quantity; or that differences in the composition of samples at each site (e.g. the proportion of high-likelihood infants or the distance families travelled to the lab) could have affected data yields. These factors are hard to disentangle, but in future analyses exploring the relation between data quantity, quality and core dependent metrics will be critical.

Some factors we were unable to standardise given available resources, like hardware used for recording. However, heterogeneity can be a strength, because measures that are going to be used in a clinical context will have to be robust to substantially more environmental variation than is common between individual babylabs. In addition, measures derived from raw trial data and that index developmental change may remain robust in the face of lower-level site differences that are consistent across trial types. Multisite studies need to consider balancing the resources put into standardisation with the desirability of testing whether metrics are robust across natural variation between sites. In the latter case, collating as much information as possible about the nature of site differences can provide valuable information about possible factors influencing the validity of any markers obtained.

Technical capacity is critically important. Our study required several technical innovations that were an essential component of success, and others that significantly reduced the investment in time required to handle the large quantities of data that we collected. Firstly, stimulus presentation can be extremely complicated for infant research due to a need

to maintain high levels of experimental control for a population that cannot be instructed, rewarded, or reasoned with. The stimuli need to be engaging for infants, and tasks must usually be split up and interspersed with each other. Most commercial stimulus presentation packages are not flexible enough for these requirements, particularly for gaze-contingent eye tracking tasks, yet custom-written scripts are orders of magnitude more complex to develop and adjust, leading to far greater potential for error in how stimuli are presented and data collected. The stimulus presentation framework we used took considerable investment to develop and test but paid dividends as we developed more tasks and added more sites to the project. Pre-processing and analysis are considerably more complex in multi-site contexts and with large samples, and a pipelined approach to stages of processing is highly recommended. Reproducibility of processing steps is vital and should be in place at an early stage to allow rapid QC and feedback to testing sites. As with stimulus presentation, automation of data processing is a costly investment at the start of a project, but the use of general purpose transformations and algorithms to handle raw data allows for code to be reused across analyses. Larger samples justify a greater degree of automation, and manual checking is often not feasible, particularly for lengthy task batteries such as we employed. However, automated processes are often a blunt tool. We recommend using automated algorithms to inspect data and highlight outliers and potentially problematic datasets for manual inspection.

Linguistic and cultural barriers

Relative to a multi-site study within a country, we were further challenged by the considerable linguistic and cultural heterogeneity within our consortium. This is particularly challenging when using stimuli with social and vocal content. Our naturalistic social videos included nursery rhymes that were selected to be relatively familiar to English-speaking infants. Using identical videos across contexts would have meant that only a proportion of the infants would be able to access the semantic content of the rhymes and would find them generally familiar at the linguistic level. Thus, we created new naturalistic social videos at every site, using local nursery rhymes that were selected to be as comparable as possible in the use of gesture, duration, and likely familiarity to the infant. At some sites the same nursery rhymes could be used (e.g. Sweden), whilst at others (e.g. Netherlands) we had to switch some of the rhymes because they were not commonly experienced. We also attempted to standardise the visual appearance, size and luminance of the movies as much as possible, whilst ensuring that their naturalistic quality was not compromised.

For the remainder of our neurocognitive tasks, standardisation was much easier than for behavioural measures. The gap, PLR, and natural scenes batteries could all be used in the exact same format at every site. This is a considerable strength of using neurocognitive assessments. In contrast, our entire behavioural protocol had to be available in the child's native language at each location. For some measures, this proved problematic (e.g. Bölte et al., 2016). For example, the Connor's is not available in some European languages, and thus these sites used the ADHD-Rating Scale for assessment of ADHD symptoms. This creates challenges in pooling this data at the analytic stage. Manufacturers commonly refuse to allow questionnaires to be translated for particular research studies or require the cost of translation to be covered by the research study but the profits from selling the measure accrue to the company. This can severely limit the applicability of standardised measures in other cultures, since the availability of translations in European languages is relatively patchy (Bölte et al., 2016). For other measures translations are available, but the data is compared to US norms (e.g. the Mullen). It is unclear whether these norms are applicable to other countries or in other languages. Indeed, a recent study indicated that the norms for the Bayley scales (a similar measure of infant development) do not operate in the same way in the Netherlands as

in the original US standardisation sample (Steenis et al., 2015). This may be one explanation for the highly variable Mullen profiles seen in our current cohorts relative to other studies. Nonetheless, the relative ease of standardising most neurocognitive measures across sites relative to behavioural or questionnaire-based measures is a considerable strength of these methodologies.

Future efforts

We are currently working on the first data release from our cohort and expect our first wave of empirical papers to be completed in 2019. We are also building our program in a range of other ways. Consortium members are beginning studies of novel groups with elevated likelihood of atypical neurodevelopment (including infants with rare genetic syndromes, premature infants; and infants with older siblings with ADHD) using our common protocol. Using identical tasks across multiple groups will allow us to ask questions about the specificity or generalisability of our early markers to other routes to ASD and related neurodevelopmental disorders. We are also working to develop consensus across our consortium on key parameters for EEG and eyetracking processing that we will cascade through other studies in our lab, including the use of several of the same experiments and principles in the Longitudinal European Autism Project (Loth et al., 2017). We hope that this will create new levels of standardisation in these fields, which have sometimes lagged behind MRI in the widespread availability of processing pipelines. We will use a similar model to expand the use of core data quality metrics in developmental eyetracking research, which are commonly omitted from many studies in this area (though see e.g. Hessels et al., 2015; Wass et al., 2014). Our employment of a consistent system for categorising reasons for data loss in EEG studies (following international guidelines developed by Webb et al., 2013) also sets a new standard for the field. Finally, our study should provide a test case for the ability to translate 'biomarkers' across differences in hardware and software. This is critical for clinical utility, but there are surprisingly few demonstrations of the limits to which insights derived from one system can be generalised. This is particularly important for biomarkers derived from machine-learning approaches, which may be very likely to overfit features of the data that are system-specific.

Conclusions

The new era of 'big science' is coming to infancy research. Within the Eurosibs consortium, we have developed both technical and conceptual approaches to dealing with the challenges this brings in the context of prospective studies of early autism. We have learned many lessons and have continued to evolve our practices to optimise the quality of data collected by our network and look forward to continuing these efforts over the next decade. Core challenges remain in achieving high comparability in behavioral data across European countries with different cultures and languages, and research remains limited by overly-prescriptive publication practices. Neurocognitive measures can be collected with high fidelity and with remarkably similar quality across sites; their great promise will be assessed once we have further knowledge of their clinical utility. Taken together, we hope that the Eurosibs study will provide a model for future data pooling efforts in early infancy research.

Acknowledgements
We would like to thank the parents and infants who participated in our research study, and the many organisations who helped with recruitment. The Eurosibs Consortium consists of: Chloe Taylor, Leila Dafner, Sarah Kalwarowsky, Nele Dewaele, Melda Arslan, Par Nystrom, Gian Candrian, Anna Malinowska, Ewa Pisula, Rafał Kawa, Maretha de Jonge, Nicolette Munsters, Lilli van Wielink, Karlijn Blommers, Declan Murphy, Grainne McAlonan. The research leading to these results received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no 115300 (EU-AIMS), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007 - 2013) and EFPIA companies' in kind contribution. This research was also supported by European Commission's Horizon 2020 Program under grant agreement no 642990 (BRAINVIEW); the Wellcome Institutional Strategic Support Fund (Birkbeck); the Research Foundation Flanders, Ghent University Special Research Fund and the Support Fund Marguerite-Marie Delacroix; the Polish National Science Centre (2012/07/B/HS6/01464); MRC Programme Grant no. G0701484 (MR/K021389/1), and the BASIS funding consortium led by Autistica. The authors declare that they have no conflicts of interest.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Bedford, R., Gliga, T., Shephard, E., Elsabbagh, M., Pickles, A., Charman, T., & Johnson, M. H. (2017). Neurocognitive and observational markers: prediction of autism spectrum disorder from infancy to mid-childhood. *Molecular autism*, *8*(1), 49.

Bedford, R., Pickles, A., Gliga, T., Elsabbagh, M., Charman, T., & Johnson, M. H. (2014). Additive effects of social and non-social attention during infancy relate to later autism spectrum disorder. *Developmental Science*, *17*(4), 612-620.

Bedford, R., Pickles, A., & Lord, C. (2016). Early gross motor skills predict the subsequent development of language in children with autism spectrum disorder. *Autism research*, *9*(9), 993-1001.

Boersma, M., Kemner, C., de Reus, M. A., Collin, G., Snijders, T. M., Hofman, D., ... & van den Heuvel, M. P. (2013). Disrupted functional brain networks in autistic toddlers. *Brain connectivity*, *3*(1), 41-49.

Bölte, S., Tomalski, P., Marschik, P. B., Berggren, S., Norberg, J., Falck-Ytter, T., ... & Kostrzewa, E. (2016). Challenges and Inequalities of Opportunities in European Psychiatry Research. *European Journal of Psychological Assessment*. doi.org/10.1027/1015-5759/a000340

Bosl, W., Tierney, A., Tager-Flusberg, H., & Nelson, C. (2011). EEG complexity as a biomarker for autism spectrum disorder risk. *BMC medicine*, *9*(1), 18.

Braukmann, R., Lloyd-Fox, S., Blasi, A., Johnson, M. H., Bekkering, H., Buitelaar, J. K., & Hunnius, S. (2017). Diminished socially selective neural processing in 5-month-old infants at high familial risk of autism. *European Journal of Neuroscience*.

Brett, D., Warnell, F., McConachie, H., & Parr, J. R. (2016). Factors affecting age at ASD diagnosis in UK: no evidence that diagnosis age has decreased between 2004 and 2014. *Journal of autism and developmental disorders*, *46*(6), 1974-1984.

Buescher, A. V., Cidav, Z., Knapp, M., & Mandell, D. S. (2014). Costs of autism spectrum disorders in the United Kingdom and the United States. *JAMA pediatrics*, *168*(8), 721-728.

Bussu, G., Jones, E. J., Charman, T., Johnson, M. H., Buitelaar, J. K., & BASIS Team. (2018). Prediction of Autism at 3 Years from Behavioural and Developmental Measures in High-Risk Infants: A Longitudinal Cross-Domain Classifier Analysis. *Journal of autism and developmental disorders*, 1-16.

Charman, T., Young, G. S., Brian, J., Carter, A., Carver, L. J., Chawarska, K., ... & Hertz-Picciotto, I. (2017). Non-ASD outcomes at 36 months in siblings at familial risk for autism spectrum disorder (ASD): A baby siblings research consortium (BSRC) study. *Autism Research*, *10*(1), 169-178.

Chawarska, K., Macari, S., & Shic, F. (2013). Decreased spontaneous attention to social scenes in 6-month-old infants later diagnosed with autism spectrum disorders. *Biological psychiatry*, *74*(3), 195-203.

Cheung, C. H. M., Bedford, R., Johnson, M. H., Charman, T., & Gliga, T. (2016). Visual search performance in infants associates with later ASD diagnosis. *Developmental cognitive neuroscience*.

Cousijn, J., Hessels, R. S., Van der Stigchel, S., & Kemner, C. (2017). Evaluation of the Psychometric Properties of the Gap-Overlap Task in 10-Month-Old Infants. *Infancy*, *22*(4), 571-579.

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., ... & Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. *Pediatrics*, *125*(1), e17-e23.

Elison, J. T., Paterson, S. J., Wolff, J. J., Reznick, J. S., Sasson, N. J., Gu, H., ... & Gerig, G. (2013). White matter microstructure and atypical visual orienting in 7-month-olds at risk for autism. *American Journal of Psychiatry*, *170*(8), 899-908.

Elsabbagh, M., Fernandes, J., Webb, S. J., Dawson, G., Charman, T., & Johnson, M. H. (2013). Disengagement of visual attention in infancy is associated with emerging autism in toddlerhood. *Biological Psychiatry*, *74*(3), 189-194.

Fletcher-Watson, S., Apicella, F., Auyeung, B., Beranova, S., Bonnet-Brilhault, F., Canal-Bedia, R., ... & Farroni, T. (2017). Attitudes of the autism community to early autism research. *Autism*, *21*(1), 61-74.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... & Lew-Williams, C. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421-435.

Gabard-Durnam, L., Tierney, A. L., Vogel-Farley, V., Tager-Flusberg, H., & Nelson, C. A. (2015). Alpha asymmetry in infants at risk for autism spectrum disorders. *Journal of autism and developmental disorders*, *45*(2), 473-480.

Gliga, T., Bedford, R., Charman, T., Johnson, M. H., Baron-Cohen, S., Bolton, P., ... & Gammer, I. (2015). Enhanced visual search in infancy predicts emerging autism symptoms. *Current Biology*, *25*(13), 1727-1730.

Gliga, T., Elsabbagh, M., Andravizou, A., & Johnson, M. (2009). Faces attract infants' attention in complex displays. *Infancy*, *14*(5), 550-562.

Green, J., Pickles, A., Pasco, G., Bedford, R., Wan, M. W., Elsabbagh, M., ... & Charman, T. (2017). Randomised trial of a parent-mediated intervention for infants at high risk for autism: longitudinal outcomes to age 3 years. *Journal of Child Psychology and Psychiatry*.

Haman, E., Łuniewska, M., Hansen, P., Simonsen, H. G., Chiat, S., Bjekić, J., ... & Gagarina, N. (2017). Noun and verb knowledge in monolingual preschool children across 17 languages:

Data from Cross-linguistic Lexical Tasks (LITMUS-CLT). *Clinical linguistics & phonetics*, *31*(11-12), 818-843.

Hazlett, H. C., Gu, H., Munsell, B. C., Kim, S. H., Styner, M., Wolff, J. J., ... & Collins, D. L. (2017). Early brain development in infants at high risk for autism spectrum disorder. *Nature*, *542*(7641), 348.

Hendry, A., Jones, E. J., Bedford, R., Gliga, T., Charman, T., & Johnson, M. H. (2018). Developmental change in look durations predicts later effortful control in toddlers at familial risk for ASD. *Journal of neurodevelopmental disorders*, *10*(1), 3.

Herlihy, L., Knoch, K., Vibert, B., & Fein, D. (2015). Parents' first concerns about toddlers with autism spectrum disorder: Effect of sibling status. *Autism*, *19*(1), 20-28.

Hessels, R. S., Hooge, I. T., & Kemner, C. (2016). An in-depth look at saccadic search in infancy. *Journal of Vision*, *16*(8), 10-10.

Hessels, R. S., Andersson, R., Hooge, I. T., Nyström, M., & Kemner, C. (2015). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, *20*(6), 601-633.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*(10-12), 1489-1506.

Johnson, M. H., Posner, M. I., & Rothbart, M. K. (1991). Components of visual orienting in early infancy: Contingency learning, anticipatory looking, and disengaging. *Journal of cognitive neuroscience*, *3*(4), 335-344.

Jones, E. J. H., Gliga, T., Bedford, R., Charman, T., & Johnson, M. H. (2014). Developmental pathways to autism: A review of prospective studies of infants at risk. *Neuroscience and Biobehavioral Reviews*.

Jones, E. J. H., Venema, K., Earl, R., Lowy, R., Barnes, K., Estes, A., ... & Webb, S. J. (2016). Reduced engagement with social stimuli in 6-month-old infants with later autism spectrum disorder: a longitudinal prospective study of infants at high familial risk. *Journal of neurodevelopmental disorders*, *8*(1), 7.

Jones, W., & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, *504*(7480), 427.

Klin, A., Shultz, S., & Jones, W. (2015). Social visual engagement in infants and toddlers with autism: early developmental transitions and a model of pathogenesis. *Neuroscience & Biobehavioral Reviews*, *50*, 189-203.

Landry, R., & Bryson, S. E. (2004). Impaired disengagement of attention in young children with autism. *Journal of Child Psychology and Psychiatry*, *45*(6), 1115-1122.

Lazenby, D. C., Sideridis, G. D., Huntington, N., Prante, M., Dale, P. S., Curtin, S., . . . Dobkins, K. (2016). Language differences at 12 months in infants who develop autism spectrum disorder. *Journal Of Autism And Developmental Disorders*, *46*(3), 899-909.

Levin, A. R., Varcin, K. J., O'Leary, H. M., Tager-Flusberg, H., & Nelson, C. A. (2017). EEG power at 3 months in infants at high familial risk for autism. *Journal of neurodevelopmental disorders*, *9*(1), 34.

Lloyd-Fox, S., Blasi, A., Elwell, C. E., Charman, T., Murphy, D., & Johnson, M. H. (2013). Reduced neural sensitivity to social stimuli in infants at risk for autism. *Proc. R. Soc. B*, *280*(1758), 20123026.

Lloyd-Fox, S., Blasi, A., Pasco, G., Gliga, T., Jones, E. J. H., Murphy, D. G. M., ... & Johnson, M. H. (2017). Cortical responses before 6 months of life associate with later autism. *European Journal of Neuroscience*.

Loth, E., Charman, T., Mason, L., Tillmann, J., Jones, E. J., Wooldridge, C., ... & Banaschewski, T. (2017). The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Molecular autism*, *8*(1), 24.

Loth, E., Garrido, L., Ahmad, J., Watson, E., Duff, A., & Duchaine, B. (2018). Facial expression recognition as a candidate marker for autism spectrum disorder: how frequent and severe are deficits?. *Molecular autism*, *9*(1), 7.

Messinger, D., Young, G. S., Ozonoff, S., Dobkins, K., Carter, A., Zwaigenbaum, L., ... & Hutman, T. (2013). Beyond autism: a baby siblings research consortium study of high-risk children at three years of age. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(3), 300-308.

Messinger, D. S., Young, G. S., Webb, S. J., Ozonoff, S., Bryson, S. E., Carter, A., ... & Dobkins, K. (2015). Early sex differences are not autism-specific: A Baby Siblings Research Consortium (BSRC) study. *Molecular autism*, *6*(1), 32.

Miller, M., Iosif, A. M., Young, G. S., Hill, M. M., & Ozonoff, S. (2016). Early detection of ADHD: insights from infant siblings of children with autism. *Journal of Clinical Child & Adolescent Psychology*, 1-8.

Mullen, E. M. (1995). *Mullen scales of early learning* (pp. 58-64). Circle Pines, MN: AGS.

Nyström, P., Gliga, T., Jobs, E. N., Gredebäck, G., Charman, T., Johnson, M. H., ... & Falck-Ytter, T. (2018). Enhanced pupillary light reflex in infancy is associated with autism diagnosis in toddlerhood. *Nature communications*, *9*(1), 1678.

Olson, J.E., Gotham, K. & Miller, F.K. (2012). *Simons Simplex Collection: ADOS and ADI-R training and reliability maintenance in multi-site phenotyping research.* Paper presented at International Meeting for Autism Research (IMFAR), Toronto, Canada

Ozonoff, S., Iosif, A. M., Baguio, F., Cook, I. C., Hill, M. M., Hutman, T., ... & Steinfeld, M. B. (2010). A prospective study of the emergence of early behavioral signs of autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*(3), 256-266.

Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., ... & Hutman, T. (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics*, *128*(3), e488-e495.

Ozonoff, S., Young, G. S., Landa, R. J., Brian, J., Bryson, S., Charman, T., ... & Zwaigenbaum, L. (2015). Diagnostic stability in young children at risk for autism spectrum disorder: a baby siblings research consortium study. *Journal of Child Psychology and Psychiatry*, *56*(9), 988-998.

Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child development*, *78*(4), 1255-1264.

Pickles, A., Le Couteur, A., Leadbitter, K., Salomone, E., Cole-Fletcher, R., Tobin, H., ... & Aldred, C. (2016). Parent-mediated social communication therapy for young children with autism (PACT): long-term follow-up of a randomised controlled trial. *The Lancet*, *388*(10059), 2501-2509.

Piven, J., & Swanson, M. R. (2017). Neurodevelopment of autism: The first three years of life. In *Autism Imaging and Devices* (pp. 53-74). CRC Press.

Righi, G., Tierney, A. L., Tager-Flusberg, H., & Nelson, C. A. (2014). Functional connectivity in the first year of life in infants at risk for autism spectrum disorder: an EEG study. *PLoS One*, *9*(8), e105176.

Salomone, E., Charman, T., McConachie, H., & Warreyn, P. (2015). Prevalence and correlates of use of complementary and alternative medicine in children with autism spectrum disorder in Europe. *European journal of pediatrics*, *174*(10), 1277-1285.

Shic, F., Macari, S., & Chawarska, K. (2014). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological psychiatry*, *75*(3), 231-237.

Sparrow, S. S., Balla, D. A., Cicchetti, D. V., Harrison, P. L., & Doll, E. A. (1984). Vineland adaptive behavior scales.

Sparrow, S. S., Cicchetti, D. V., Balla, D. A., & Doll, E. A. (2005). *Vineland adaptive behavior scales: Survey forms manual*. American Guidance Service.

Steenis, L. J., Verhoeven, M., Hessen, D. J., & Van Baar, A. L. (2015). Performance of Dutch children on the Bayley III: a comparison study of US and Dutch norms. *PloS one*, *10*(8), e0132871.

Szatmari, P., Chawarska, K., Dawson, G., Georgiades, S., Landa, R., Lord, C., ... & Halladay, A. (2016). Prospective longitudinal studies of infant siblings of children with autism: lessons learned and future directions. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(3), 179-187.

Tick, B., Bolton, P., Happé, F., Rutter, M., & Rijsdijk, F. (2016). Heritability of autism spectrum disorders: a meta-analysis of twin studies. *Journal of Child Psychology and Psychiatry*, *57*(5), 585-595.

Tierney, A. L., Gabard-Durnam, L., Vogel-Farley, V., Tager-Flusberg, H., & Nelson, C. A. (2012). Developmental trajectories of resting EEG power: an endophenotype of autism spectrum disorder. *PloS one*, *7*(6), e39127.

Van Den Boomen, C., de Graaff, J. C., de Jong, T. P., Kalkman, C. J., & Kemner, C. (2013). General anesthesia as a possible GABAergic modulator affects visual processing in children. *Frontiers in cellular neuroscience*, *7*, 42.

Van Rijn, A. M., Peper, A., & Grimbergen, C. A. (1990). High-quality recording of bioelectric events. *Medical and Biological Engineering and Computing*, *28*(5), 389-397.

Vlamings, P. H., Jonkman, L. M., & Kemner, C. (2010a). An Eye for Detail: An Event-Related Potential Study of the Rapid Processing of Fearful Facial Expressions in Children. *Child development*, *81*(4), 1304-1319.

Vlamings, P. H. J. M., Jonkman, L. M., van Daalen, E., van der Gaag, R. J., & Kemner, C. (2010b). Basic abnormalities in visual processing affect face processing at an early age in autism spectrum disorder. *Biological psychiatry*, *68*(12), 1107-1113.

Wass, S. V., Forssman, L., & Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy*, *19*(5), 427-460.

Wass, S. V., & Smith, T. J. (2014). Individual differences in infant oculomotor behavior during the viewing of complex naturalistic scenes. *Infancy*, *19*(4), 352-384.

Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229-250.

Webb, S.J., Bernier, R., Henderson, H.A., Johnson, M.H., Jones, E.J.H., Lerner, M.D., McPartland, J., Nelson, C.A., Rojas, D.C., Townsend, J., & Westerfield, M. (2013). Guidelines and best practices for electrophysiological data collection, analysis and reporting in autism. *Journal of Autism and Developmental Disorders, 45(2),* 425-443.

Wolff, J. J., Gerig, G., Lewis, J. D., Soda, T., Styner, M. A., Vachet, C., ... & Hazlett, H. C. (2015). Altered corpus callosum morphology associated with autism over the first 2 years of life. *Brain*, *138*(7), 2046-2058.

Woolfenden, S., Sarkozy, V., Ridley, G., & Williams, K. (2012). A systematic review of the diagnostic stability of autism spectrum disorder. *Research in Autism Spectrum Disorders*, *6*(1), 345-354.

Zwaigenbaum, L., Bryson, S., Rogers, T., Roberts, W., Brian, J., & Szatmari, P. (2005). Behavioral manifestations of autism in the first year of life. *International journal of developmental neuroscience*, *23*(2-3), 143-152.

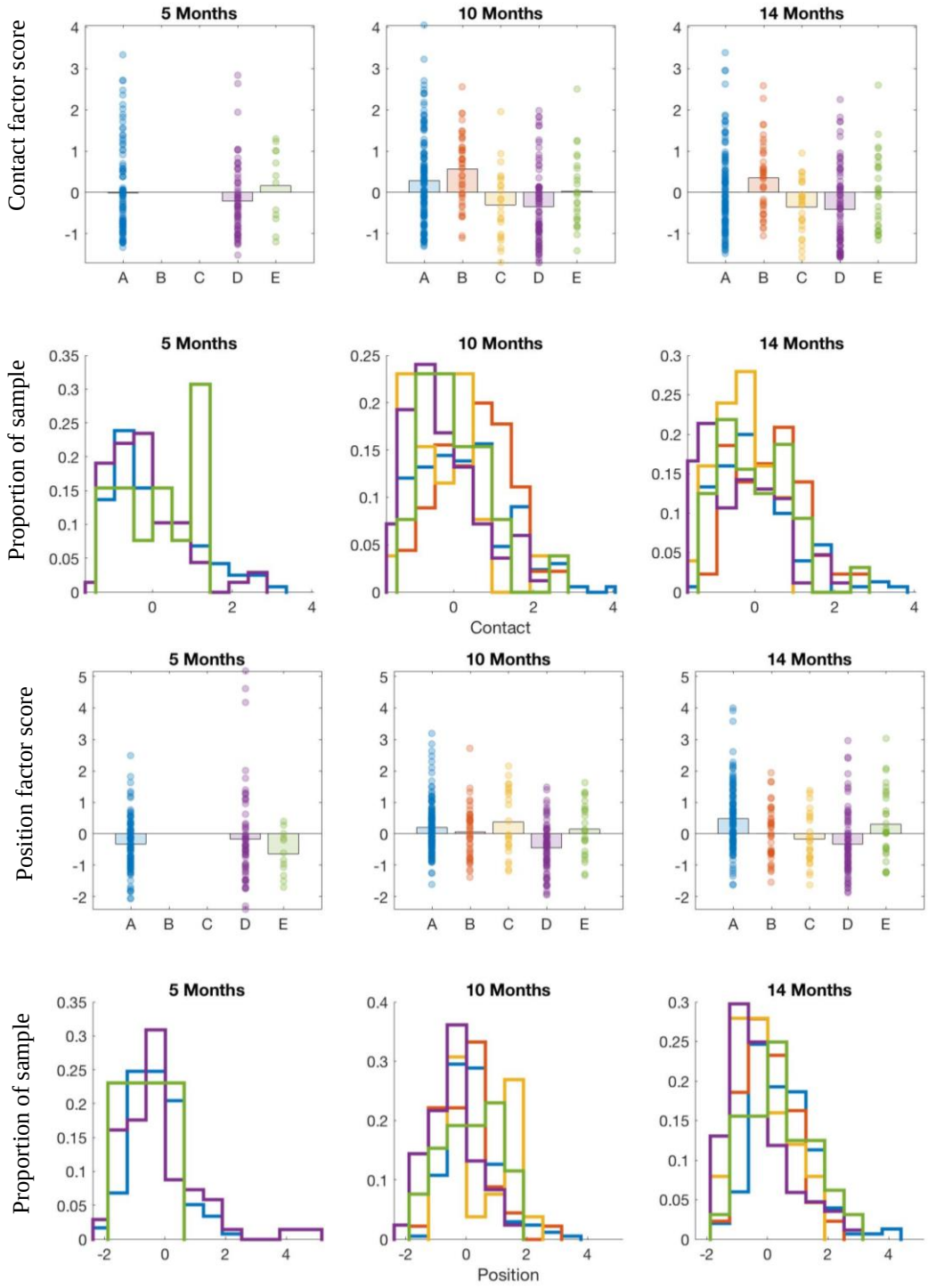

Figure 1: Eye tracking data quality metrics by age and site. Contact is a factor primarily composed of posthoc drift, proportion samples lost, and flicker ratio. Position is a factor primarily composed of variability in distance from the screen and track box. The bar charts represent mean and individual data; the histograms depict the proportion of children at each site within each bin.

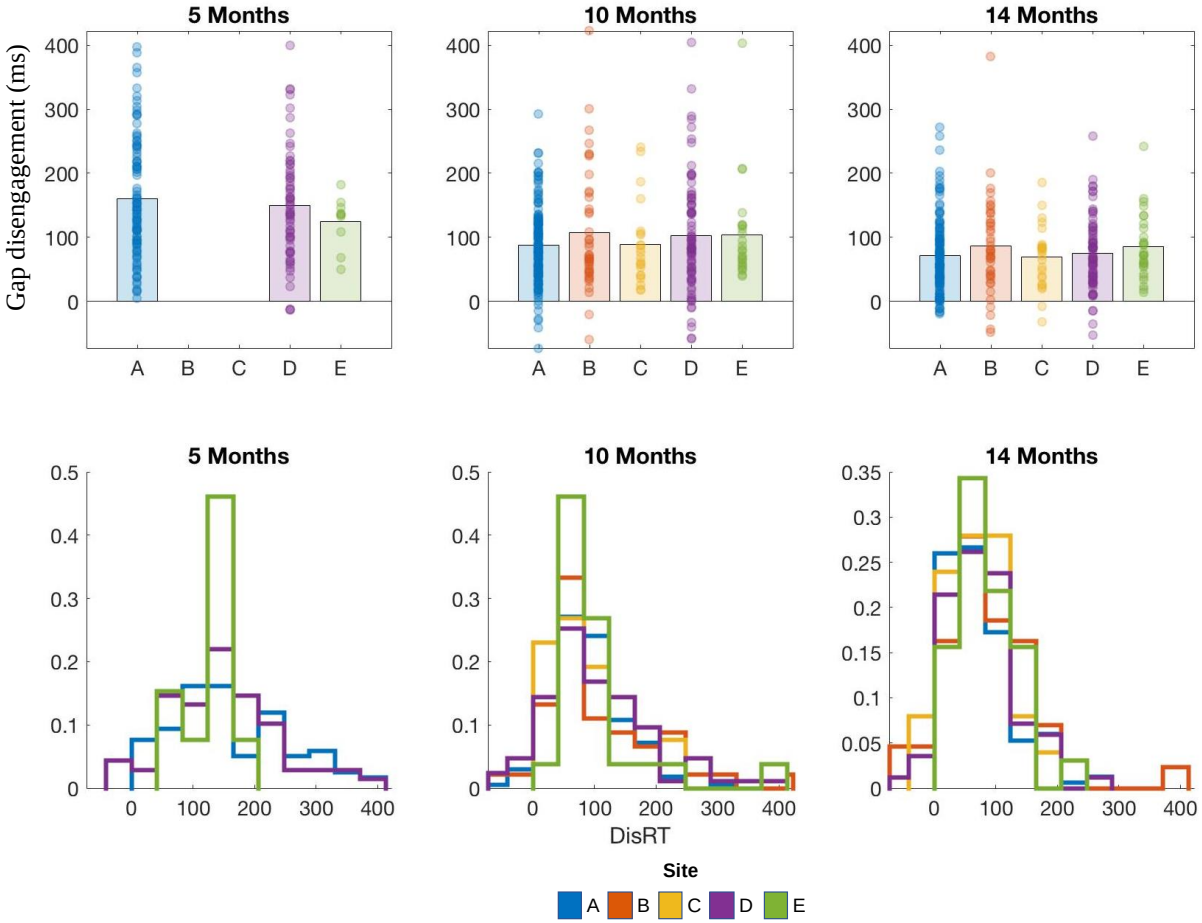

Figure 2: Gap disengagement scores by age and site. The bar charts represent mean and individual data; the histograms depicts the proportion of children at each site within each bin.

Table 1a: Participant gender balance and likelihood group for each site (%). HL = High Likelihood; LL = Low Likelihood.

	HL-ASD			*LL*		
Site	*Male*	*Female*	*Total*	*Male*	*Female*	*Total*
A (129)	43%	35%	78%	14%	8%	23%
B (51)	35%	27%	62%	22%	16%	38%
C (25)	36%	24%	60%	28%	12%	40%
D (72)	32%	38%	70%	18%	12%	30%
E (100)	29%	26%	55%	24%	21%	45%

Table 1b: Age in days by site and time-point

	5 months	*10 months*	*14 months*
A	177.0 (18.2) 117-209	319.9 (15.0) 287-358	450.7 (19.0) 406-516
B	160.5 (16.7) 141-198	307.8 (16.7) 279-370	429.9 (18.8) 371-483
C		314.9 (16.9) 271-345	433.9 (15.4) 406-464
D	160.9 (15.2) 119-207	311.0 (13.4) 276-364	433.5 (17.8) 391-478
E		314.3 (17.8) 276-353	437.2 (21.4) 363-483
Effect of site on age	$F(2,189) = 21.7$, $p < 0.001$, $\eta p^2 = 0.19$	$F(4,374) = 7.66$, $p < 0.001$, $\eta p^2 = 0.076$	$F(4,321) = 19.6$, $p < 0.001$, $\eta p^2 = 0.16$

Table 2: Maternal education across sites.

Maternal education (%)				
Site	Primary	Secondary	Tertiary Undergraduate	Tertiary Postgraduate
A	0.8%	26.4%	41.6%	25.6%
B	2%	33.3%	33.3%	17.5%
C	0%	28%	24%	36%
D	2.8%	18.1%	9.7%	45.8%
E	0%	18%	38%	43%

Table 3: Site effects on the Mullen Scales of Early Learning, computed on domain standard scores. Darker colours represent higher effect sizes. Numbers in the first column represent the numbers of infants at each site who provided data for these analyses.

	Mullen Scales of Early Learning				
	Gross motor	Visual reception	Fine motor	Rec. Lang.	Exp. Lang.
5 months A=91 B=36 D=65	$F(2, 192) = 2.05$, $p = 0.13$, $\eta p^2 = 0.021$	$F(2, 192) = 6.69$, $p = 0.002$, $\eta p^2 = 0.066$	$F(2, 192) = 5.54$, $p = 0.005$, $\eta p^2 = 0.055$	$F(2, 192) = 29.58$, $p < 0.001$, $\eta p^2 = 0.24$	$F(2, 192) = 21.52$, $p < 0.001$, $\eta p^2 = 0.185$
10 months A=119 B=46 C=14 D=87 E=61	$F(2, 327) = 7.13$, $p < 0.001$, $\eta p^2 = 0.081$	$F(2, 327) = 10.99$, $p < 0.001$, $\eta p^2 = 0.12$	$F(2, 327) = 12.16$, $p < 0.001$, $\eta p^2 = 0.13$	$F(2, 327) = 23.99$, $p < 0.001$, $\eta p^2 = 0.23$	$F(2, 356) = 34.78$, $p < 0.001$, $\eta p^2 = 0.30$
14 months A=103 B=41 D=91 E=60	$F(2, 295) = 10.36$, $p < 0.001$, $\eta p^2 = 0.096$	$F(2, 295) = 58.73$, $p < 0.001$, $\eta p^2 = 0.38$	$F(2, 295) = 22.57$, $p < 0.001$, $\eta p^2 = 0.19$	$F(2, 295) = 23.51$, $p < 0.001$, $\eta p^2 = 0.195$	$F(2, 295) = 28.22$, $p < 0.001$, $\eta p^2 = 0.23$

Table 4: Site effects on the Vineland Adaptive Behaviour Scales, computed on domain standard scores. Darker colours represent higher effect sizes. Numbers in the first column represent the numbers of infants at each site who provided data for these analyses.

	Vineland Adaptive Behavioural Scales			
	Communication	Socialisation	Motor Score	Daily Living
5 months A=78 B=32 D=63	$F(2, 172) = 0.72$, $p = 0.49$, $\eta p^2 = 0.008$	$F(2, 172) = 0.45$, $p = 0.64$, $\eta p^2 = 0.005$	$F(2, 172) = 0.057$, $p = 0.94$, $\eta p^2 = 0.001$	$F(2, 172) = 12.59$, $p < 0.001$, $\eta p^2 = 0.13$
10 months A=96 B=40 D=136	$F(2,272) = 3.74$, $p = 0.025$, $\eta p^2 = 0.027$	$F(2,272) = 7.36$, $p = 0.001$, $\eta p^2 = 0.052$	$F(2,272) = 21.76$, $p < 0.001$, $\eta p^2 = 0.14$	$F(2,272) = 21.97$, $p < 0.001$, $\eta p^2 = 0.14$
14 months A=65 B=28 D=117	$F(2,210) = 3.48$, $p = 0.033$, $\eta p^2 = 0.032$	$F(2,210) = 5.71$, $p = 0.004$, $\eta p^2 = 0.052$	$F(2,210) = 23.93$, $p < 0.001$, $\eta p^2 = 0.19$	$F(2,210) = 20.96$, $p < 0.001$, $\eta p^2 = 0.17$

Table 5: Association between Mullen and Vineland domain scores over time. Data reported are ANCOVAs with 5/10 month Mullen domain scores as the covariates, site as an independent variable, and 10/14-month Mullen domain scores as the outcome variables. Colours represent effect sizes, with darker shades being larger.

	5 to 10 months		**10 to 14 months**	
	Predictor	**Predictor x Site**	**Predictor**	**Predictor x Site**
Mullen				
Gross Motor	$F(1,175) = 25.87$, $p < 0.001$, $\eta p^2 = 0.13$	$F(2,175) = 0.57$, $p = 0.84$, $\eta p^2 = 0.002$	$F(1,265) = 107.0$, $p < 0.001$, $\eta p^2 = 0.29$	$F(3,265) = 2.23$, $p = 0.085$, $\eta p^2 = 0.025$
Visual Reception	$F(1,175) = 8.01$, $p = 0.005$, $\eta p^2 = 0.045$	$F(2,175) = 0.41$, $p = 0.67$, $\eta p^2 = 0.005$	$F(1, 310) = 22.31$, $p < 0.001$; $\eta p^2 = 0.07$	$F(3,310) = 0.29$, $p = 0.83$, $\eta p^2 = 0.003$
Fine Motor	$F(1,175) = 17.73$, $p < 0.001$, $\eta p^2 = 0.095$	$F(2,175) = 1.42$, $p = 0.25$, $\eta p^2 = 0.016$	$F(1,309) = 28.19$, $p < 0.001$, $\eta p^2 = 0.09$	$F(3,309) = 1.34$, $p = 0.26$, $\eta p^2 = 0.013$
Rec. Lang	$F(1,175) = 11.51$, $p = 0.001$, $\eta p^2 = 0.06$	$F(2,175) = 1.07$, $p = 0.35$, $\eta p^2 = 0.013$	$F(1,310) = 31.9$, $p < 0.001$, $\eta p^2 = 0.096$	$F(3,310) = 0.093$, $p = 0.96$, $\eta p^2 = 0.001$
Exp. Lang	$F(1,174) = 3.68$, $p = 0.06$, $\eta p^2 = 0.021$	$F(2,174) = 1.51$, $p = 0.22$, $\eta p^2 = 0.018$	$F(1,311) = 42.20$, $p < 0.001$, $\eta p^2 = 0.12$	$F(3,311) = 1.12$, $p = 0.34$, $\eta p^2 = 0.011$
Vineland				
Communication	$F(1,151) = 16.7$, $p < 0.001$, $\eta p^2 = 0.10$	$F(2,151) = 1.87$, $p = 0.16$, $\eta p^2 = 0.025$	$F(1,207) = 51.3$, $p < 0.001$, $\eta p^2 = 0.2$	$F(2,207) = 1.13$, $p = 0.33$, $\eta p^2 = 0.011$
Socialisation	$F(1,152) = 9.1$, $p = 0.003$, $\eta p^2 = 0.06$	$F(2, 152) = 0.92$, $p = 0.4$, $\eta p^2 = 0.012$	$F(1,203) = 16.42$, $p < 0.001$, $\eta p^2 = 0.077$	$F(2,203) = 4.34$, $p = 0.014$, $\eta p^2 = 0.042$

Motor	$F(1,149) = 5.52$, $p = 0.02$, $\eta p^2 = 0.037$	$F(2,149) = 0.12$, $p = 0.89$, $\eta p^2 = 0.002$	$F(1,206) = 60.86$, $p < 0.001$, $\eta p^2 = 0.23$	$F(2,206) = 2.54$, $p = 0.08$, $\eta p^2 = 0.025$
Daily Living	$F(1,157) = 5.80$, $p = 0.017$, $\eta p^2 = 0.037$	$F(2,157) = 0.22$, $p = 0.81$, $\eta p^2 = 0.003$	$F(1,205) = 21.94$, $p < 0.001$, $\eta p^2 = 0.099$	$F(2,205) = 0.20$, $p = 0.82$, $\eta p^2 = 0.002$

Table 6: Validity and quantity of EEG data across sites where pre-processing available.

	5 months		*10 months*		*14 months*	
	% Valid	*N trials watched*	*% Valid*	*N trials watched*	*% Valid*	*N trials watched*
A (n=89/116/ 108)	93%	5.7 (.1) *2-6*	84%	5.2 (.1) *1-6*	70%	5.3 (.14) *2-6*
B (n=0/46/39)			89%	5.4 (.16) 2-6	74%	6 (0) 6-6
C (n=0/24/24)			100%	5.6 (.16) *2-6*	88%	5.2 (.25) 2-6
D (n=66)	86%	5.3 (.13) *3-6*				
E						

Table 7: Core eyetracking matrices over time. Data reported are ANCOVAs with 5/10-month eyetracking scores as the covariates, site as an independent variable, and 10/14-month eyetracking scores as the outcome variables. Colours represent effect sizes, with darker shades being larger.

	5 to 10 months		**10 to 14 months**	
	Predictor	**Predictor x Site**	**Predictor**	**Predictor x Site**
Eyetracking variable				
Gap overall trial N	$F(107) = 5.60$, $p = .02$, $\eta p^2 = .05$	$F(107) = 1.11$, $p = .30$, $\eta p^2 = .011$	$F(191) = 2.87$, $p < .092$, $\eta p^2 = .015$	$F(191) = 1.42$, $p < .24$, $\eta p^2 = .023$
With Contact at the two ages covaried	$F(89) = 7.92$, $p = .006$, $\eta p^2 = .087$	$F(89) = 6.92$, $p = .01$, $\eta p^2 = 0.077$ (stronger site effects at Site E)	$F(164) = 3.22$, $p = .075$, $\eta p^2 = .021$	$F(164) = 1.54$, $p = .21$, $\eta p^2 = .029$
Baseline SRT	$F(90) = 8.52$, $p = 0.004$, $\eta p^2 = .09$	$F(90) = 1.03$, $p = .31$, $\eta p^2 = .012$	$F(175) = 53.76$, $p < .001$, $\eta p^2 = .24$	$F(175) = 1.27$, $p = .29$, $\eta p^2 = .022$
With Contact at the two ages covaried	$F(77) = 9.95$, $p = 0.002$, $\eta p^2 = .12$	$F(77) = 1.08$, $p = .30$, $\eta p^2 = .015$	$F(156) = 60.06$, $p < .001$, $\eta p^2 = .29$	$F(156) = 1.72$, $p = .17$, $\eta p^2 = .034$
Disengagement	$F(90) = .33$, $p = .57$, $\eta p^2 = .004$	$F(88) = .003$, $p = .96$, $\eta p^2 < .001$	$F(175) = 7.54$, $p = .007$, $\eta p^2 = .043$	$F(167) = 1.01$, $p = .393$, $\eta p^2 = .018$
With Contact at the two ages covaried	$F(77) = .37$, $p = .54$, $\eta p^2 = .005$	$F(77) = .005$, $p = .95$, $\eta p^2 < .001$	$F(156) = 10.19$, $p = .002$, $\eta p^2 = .065$	$F(156) = 1.46$, $p = .23$, $\eta p^2 = .029$
Contact	$F(94) = 7.91$, $p = .006$, $\eta p^2 = .081$	$F(94) = .335$, $p = .56$, $\eta p^2 = .004$	$F(176) = 20.46$, $p < 0.001$, $\eta p^2 = .11$	$F(176) = 4.65$, $p = .004$, $\eta p^2 = .077$
Position	$F(94) = .85$, $p = .36$, $\eta p^2 = .009$	$F(94) = .32$, $p = .58$, $\eta p^2 = .003$	$F(175) = .66$, $p = .42$, $\eta p^2 = .004$	$F(175) = .98$, $p = .40$, $\eta p^2 = .017$